LSAView: A Tool for Visual Exploration of Latent Semantic Modeling

Patricia Crossno, Daniel Dunlavy, Timothy Shead

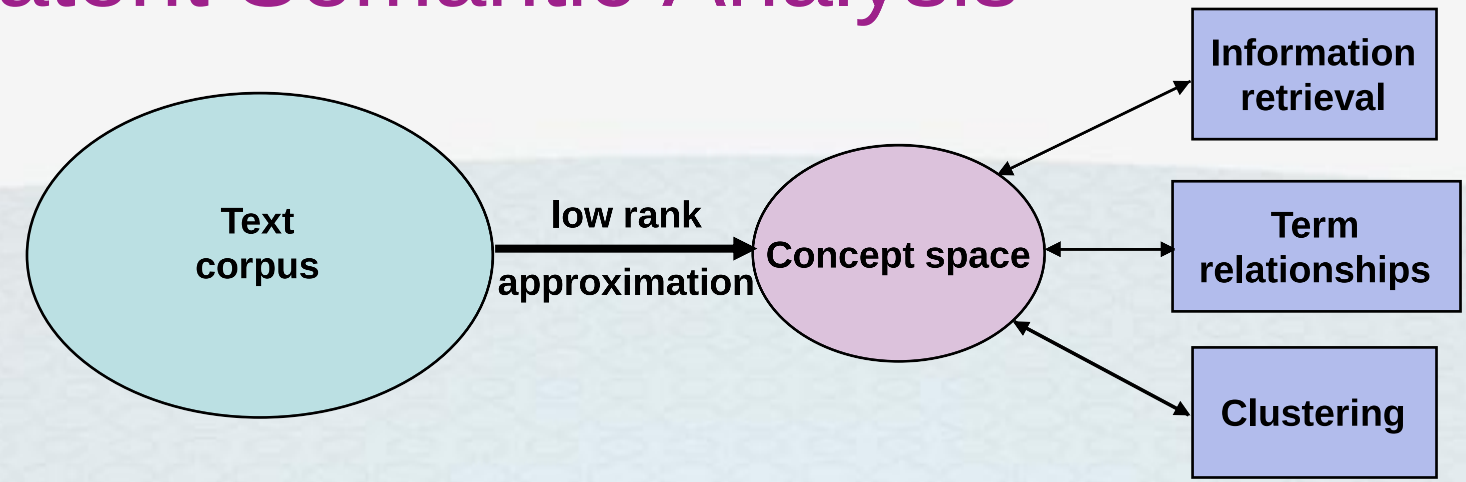
Sandia National Laboratories

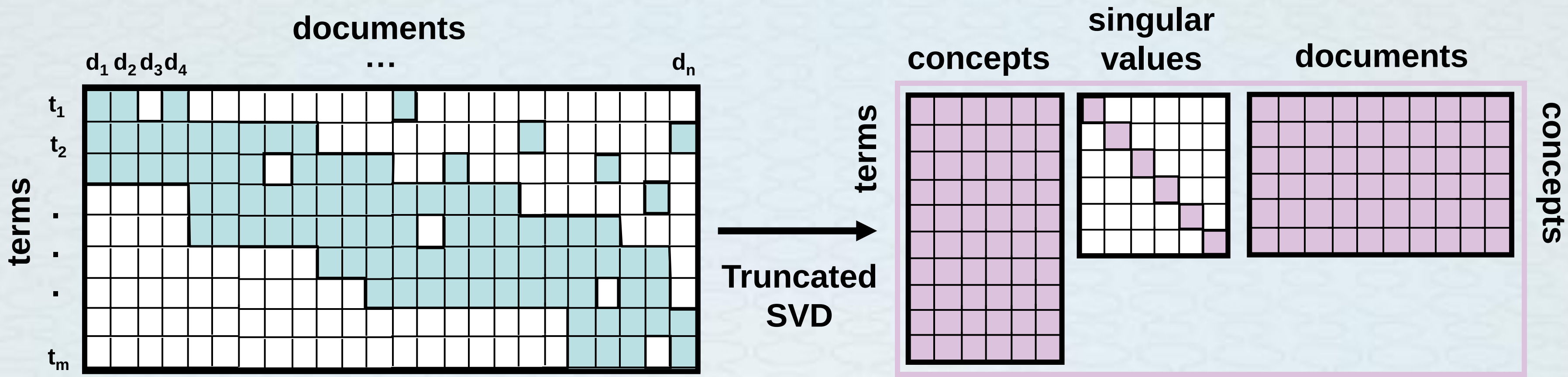
Sandia is a multi-program laboratory operated by Sandia Corporation, a Lockheed Martin Company,
for the United States Department of Energy under contract DE-AC04-94AL85000.

Overview

- Latent Semantic Analysis
- Motivation
- Analysis of Algorithmic Choices
- LSAView
- Case Studies
  - Rank Selection
  - Singular Value Scaling
- Conclusions

Latent Semantic Analysis

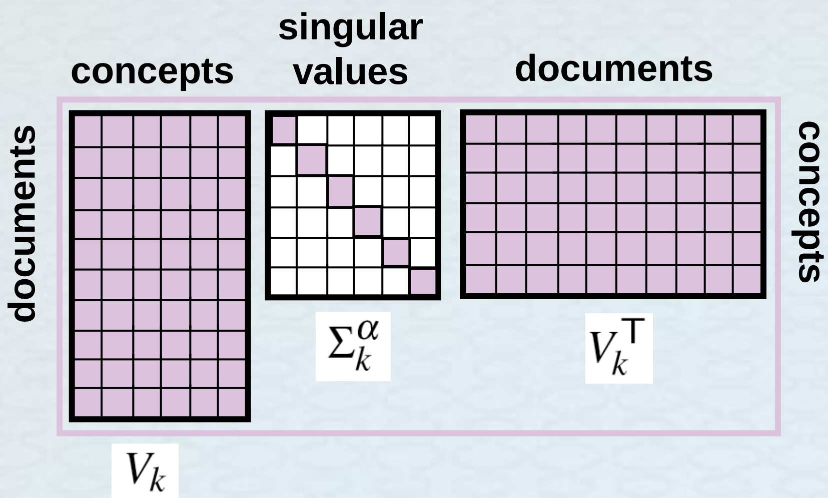




$$A \approx U_k \Sigma_k V_k^{\mathsf{T}}$$

Document Similarity Graphs

Document similarity matrix = $V_k \Sigma_k^{\alpha} V_k^{\mathsf{T}}$



$$\xrightarrow[\text{threshold}]{\text{rank} = k} \begin{bmatrix} i & j & w_{ij} \\ i & j & w_{ij} \\ \vdots & & \\ i & j & w_{ij} \end{bmatrix}$$

sparse coordinate format

Use Cosine Similarities

$$e_{ij}(k) = \frac{\langle v_k^i \Sigma_k, v_k^j \Sigma_k \rangle}{\|v_k^i \Sigma_k\|_2 \, \|v_k^j \Sigma_k\|_2}$$

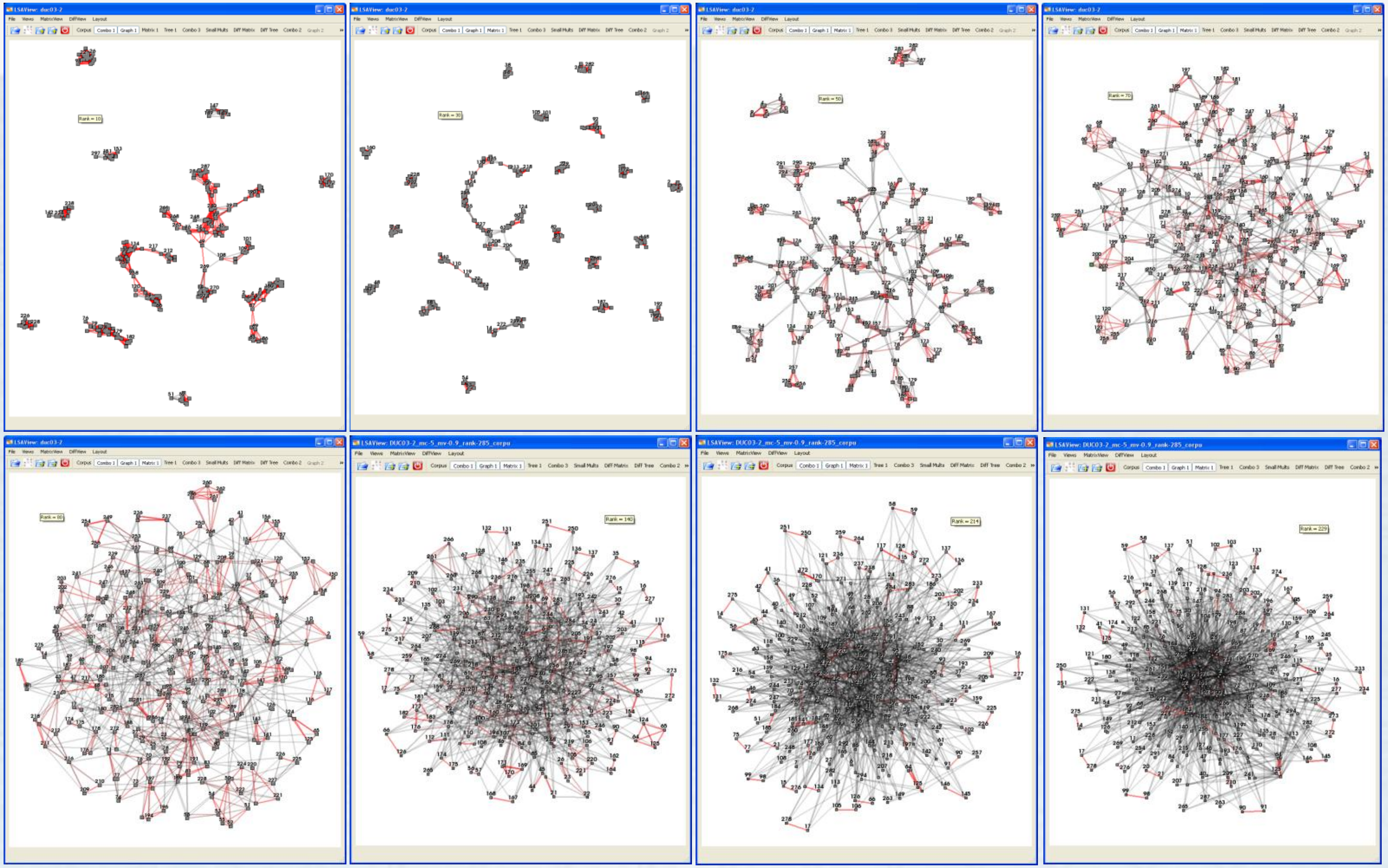
Document similarity graph

- Each document is a vertex
- Each row defines an edge

VAST 09

Motivation:

Algorithmic Parameter Choices Impact Models

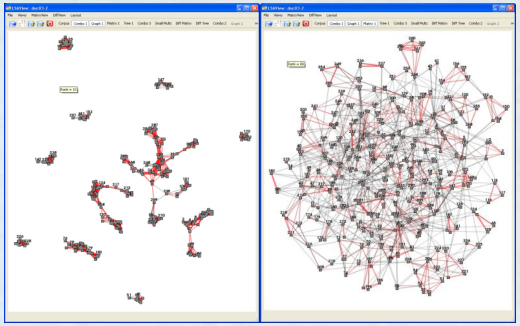


Which rank to use?

Analysis of Algorithmic Choices

Focus on impacts from:

- Rank (number of concepts)
  - Find sweet spot between extremes
- Similarity computation
  - Singular value scaling



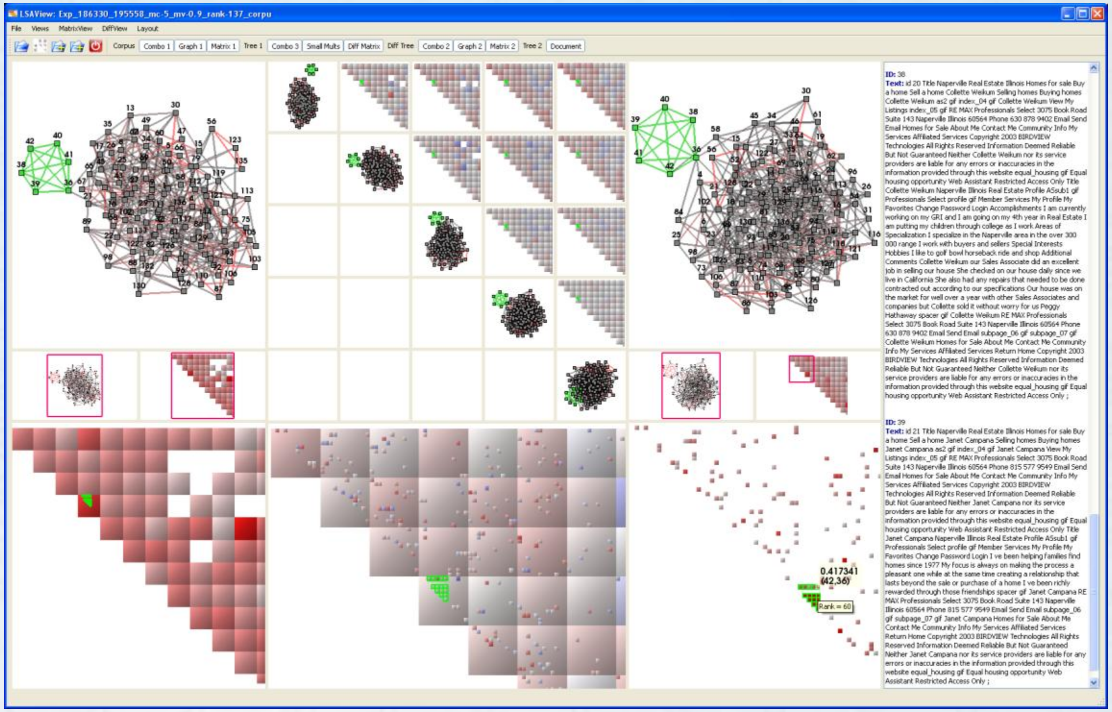
How to visualize model impacts?

- Conceptual groupings
  - Document layout
  - Changes in link strength between documents
- Significance of changes in edge weights
  - Large changes not necessarily significant
  - Statistical inference tests

VAST 09

LSAView

- Compares models
- Explores impacts of parameter choices
- Uses statistical inference to highlight model differences
- Built using open source VTK/Titan Informatics Toolkit
- Views
  - Graph
  - Matrix
  - You Are Here
  - Small Multiples
  - Document



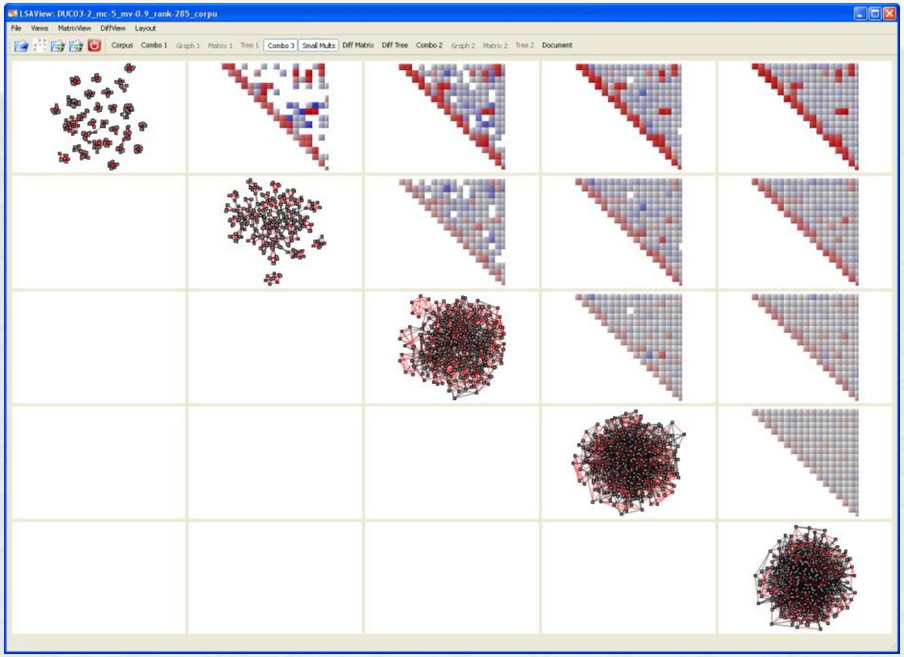
Rank Selection Case Study

- DUC data
  - 2003 Document Understanding Conference (DUC)
  - 298 newswire documents for summarization evaluation
  - Documents in 30 clusters
  - ~10 documents per cluster on a particular topic or event
  - http://www-nlpir.nist.gov/projects/duc/data.html
- Rank = k (SVD truncation)

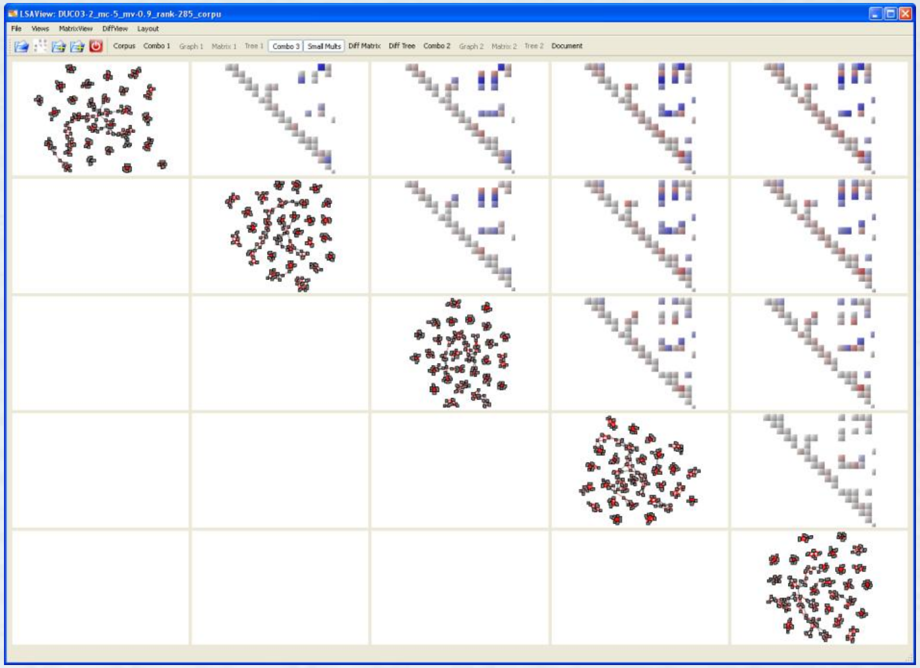
$$A \approx U_k \Sigma_k V_k^{\mathsf{T}}$$

- Iterative Approach
  - Identify range of potential ranks – *Small Multiples View*
  - Compare ranks – *Graph, Matrix, and Data Table Views*
  - Validate rank – *Document View*

Small Multiples: Narrow Range of Ranks



Ranks k = 20, 50, 80, 11, 140

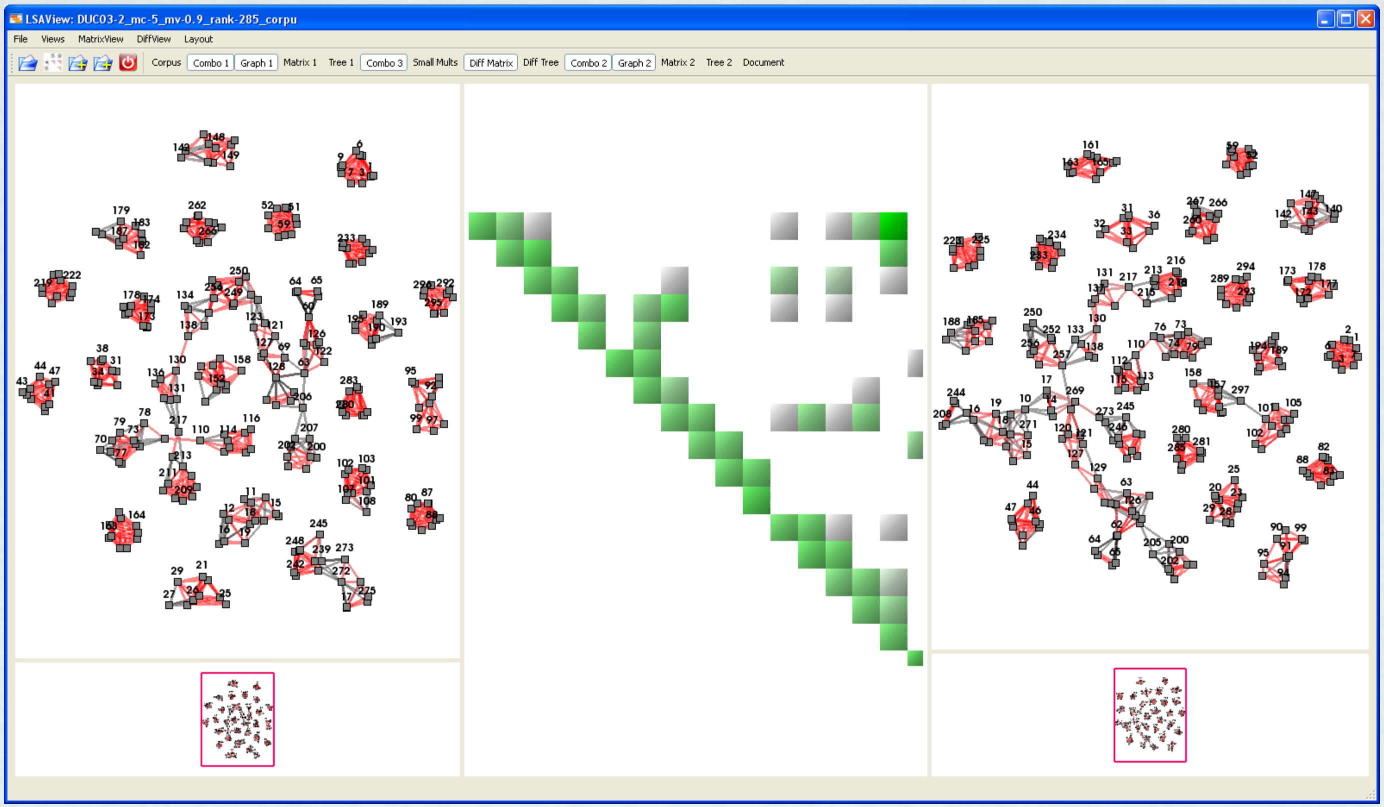


Ranks k = 28, 29, 30, 31, 32

Two-sample *t* Statistics

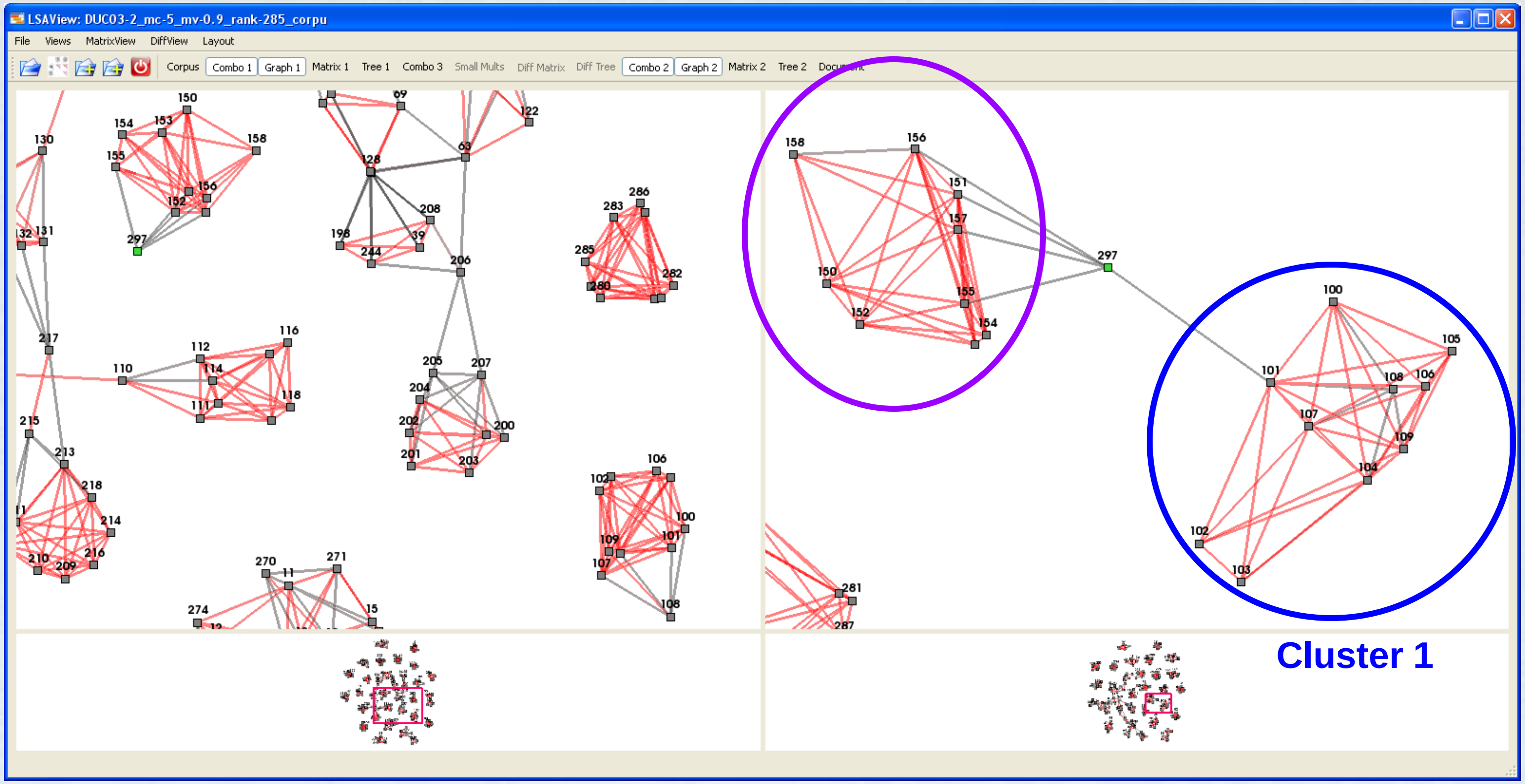
$$t_{ij}^{(2)} = \frac{\bar{e}_{ij}(k_1,\alpha,n_1) - \bar{e}_{ij}(k_2,\alpha,n_2)}{\sqrt{\frac{[s_{ij}(k_1,\alpha,n_1)]^2}{n_1} + \frac{[s_{ij}(k_2,\alpha,n_2)]^2}{n_2}}}$$

- Identify anomalous edge weights between 2 graphs
- Most significant differences in bright green



Anomalous Links to Document 297

Cluster 2



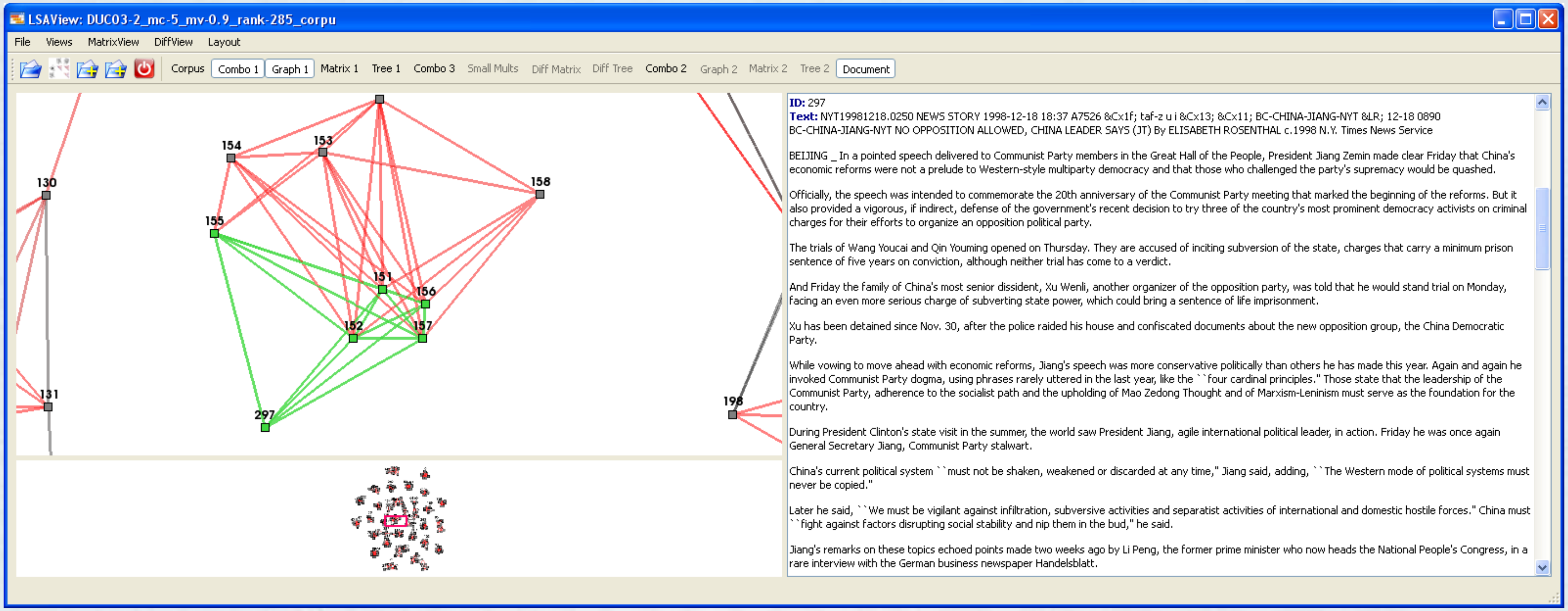
Rank 30

Rank 32

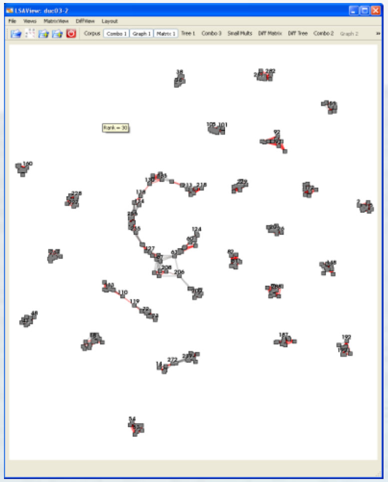
VAST 09

Manual Inspection

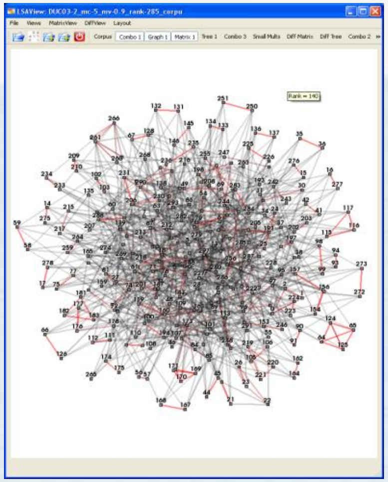
- Document 297 – Chinese policy on separatists
- Cluster 1 topic – trial of 3 Chinese separatists
- Cluster 2 topic – Russian policy on Chechnyan separatists
- Policy theme best match for 297, conclude Rank 30 best



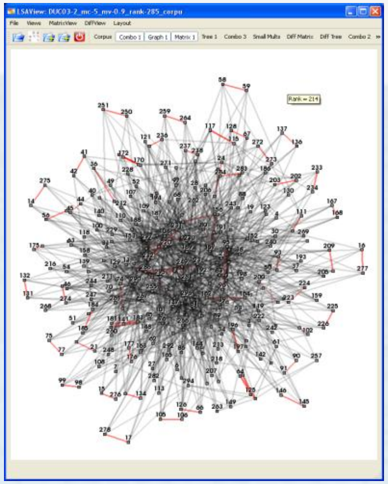
Comparison to Automated Methods



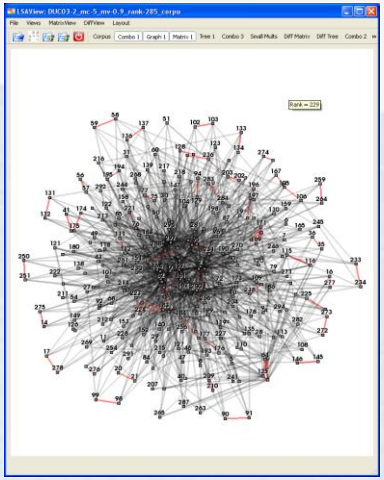
LSAView
Rank 30
Variance 40.59



Leave-1-Out
Cross Validation
Rank 140
Variance 80.72



95% Variance
Rank 214
Variance 95.12



20-group (fold)
Cross Validation
Rank 229
Variance 97.27

- Automated rank selection methods select ranks
  - Robust to noise
  - Accounting for variance in data
- LSAView selects on impact to text analysis tasks

Singular Value Scaling Case Study

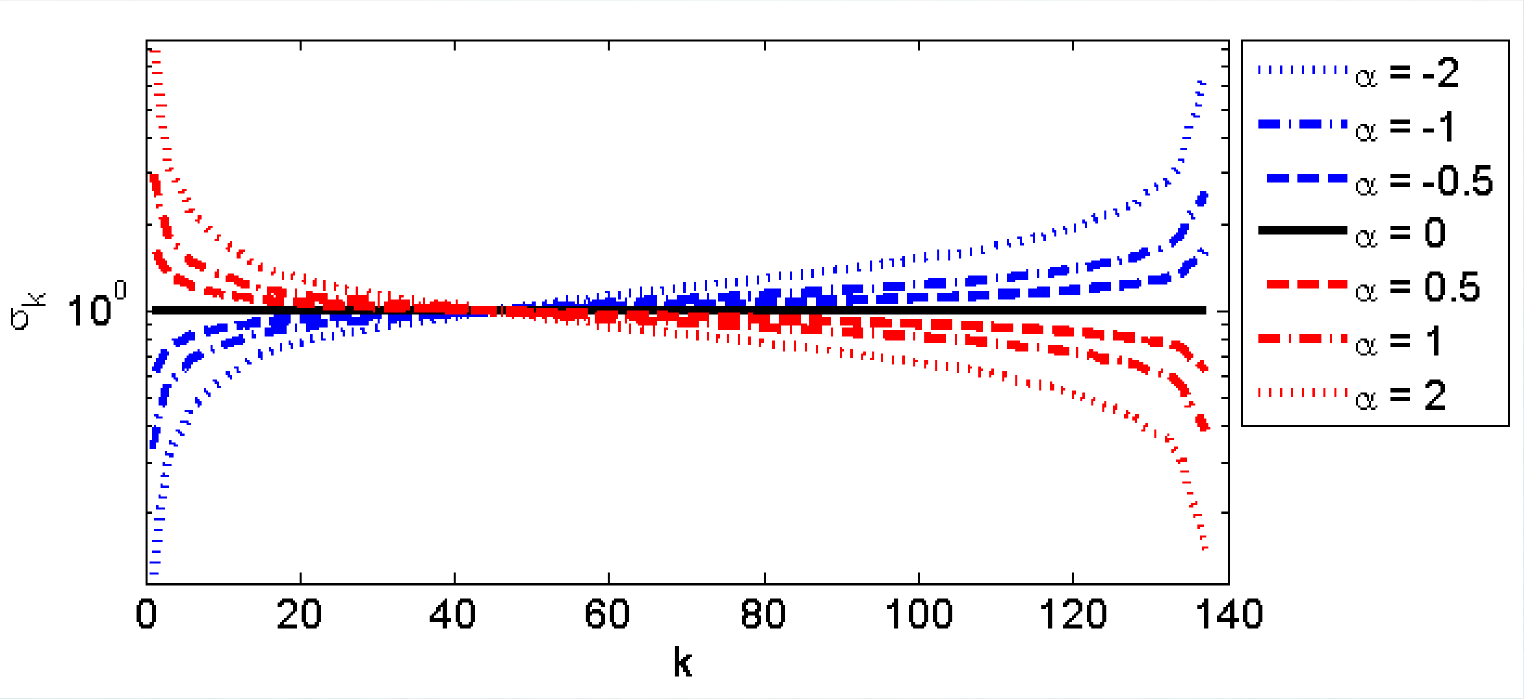
- TechTC data
  - Subset of TechTC-100 Test Collection
  - 150 html documents partitioned into 2 clusters
  - http://techtc.cs.technion.ac.il/techtc100/techtc100.html
- Singular Value Scaling = α

$$e_{ij}(k,\alpha) = \frac{\langle v_k^i \Sigma_k^{\alpha/2}, v_k^j \Sigma_k^{\alpha/2} \rangle}{\|v_k^i \Sigma_k^{\alpha/2}\|_2 \, \|v_k^j \Sigma_k^{\alpha/2}\|_2}$$

- Complicated by rank selection
- Inspect scaled singular values for α vs. k

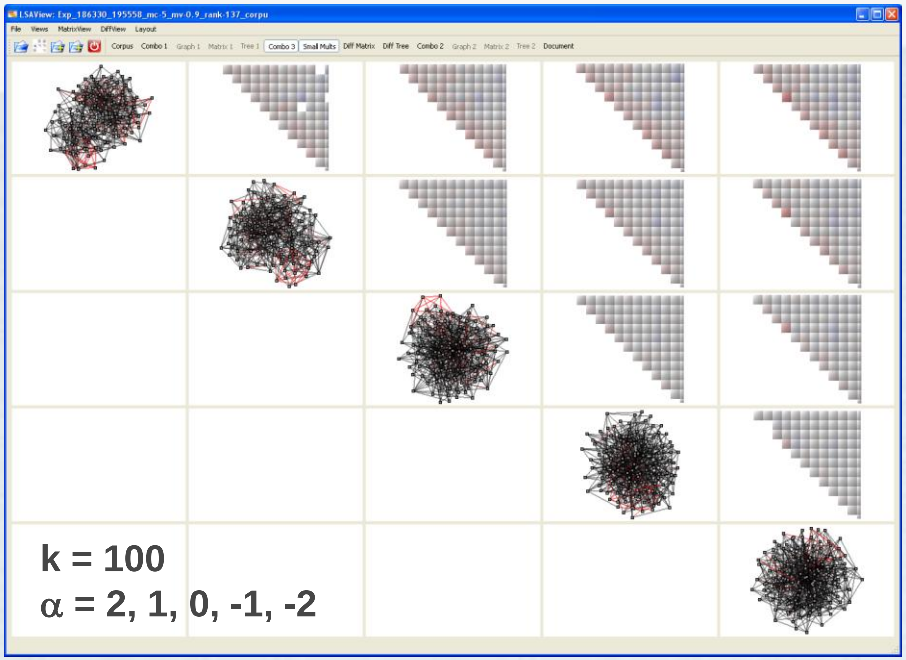
Inspect Singular Values Scaled by α

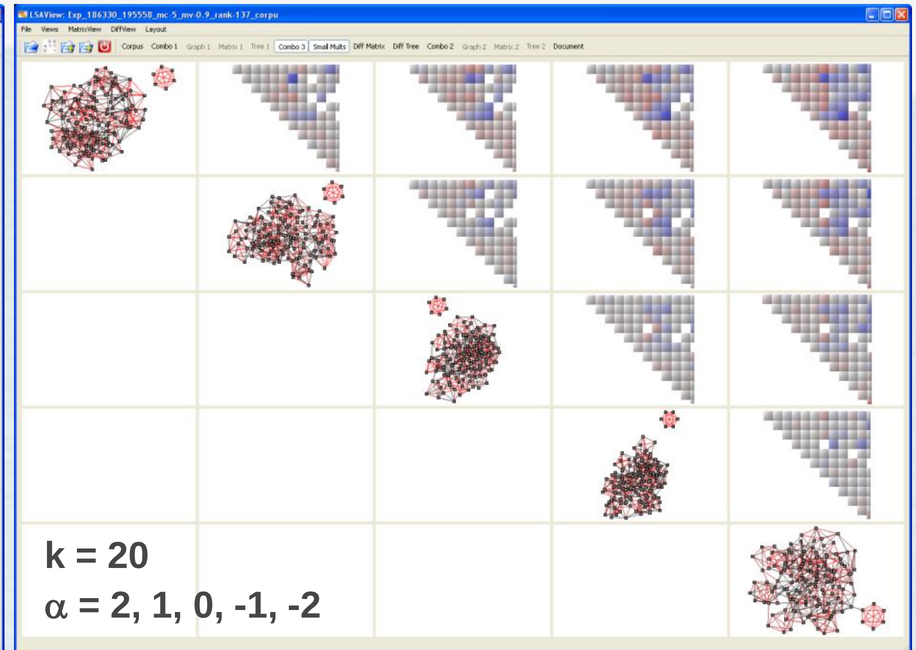
- Original singular values correspond to $\alpha = 2$
- For all α, values trend toward 0 for k < 45
- For k > 45, inverted scalings amplify noise



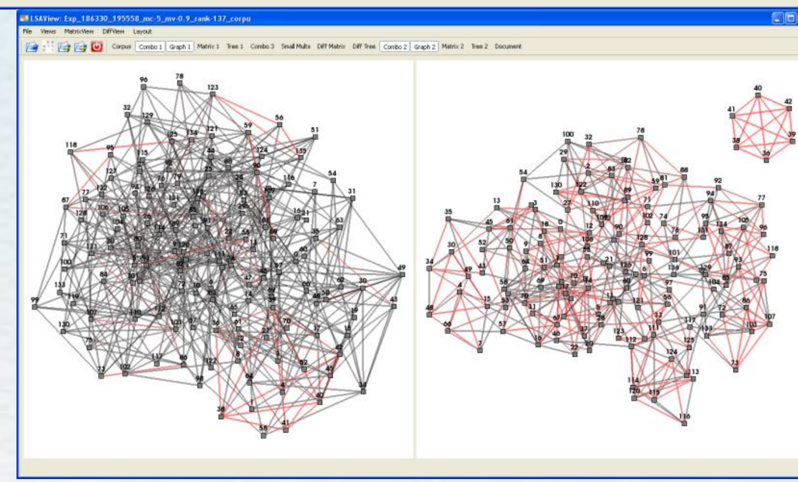
VAST 09

Small Multiples k > 45 vs k < 45

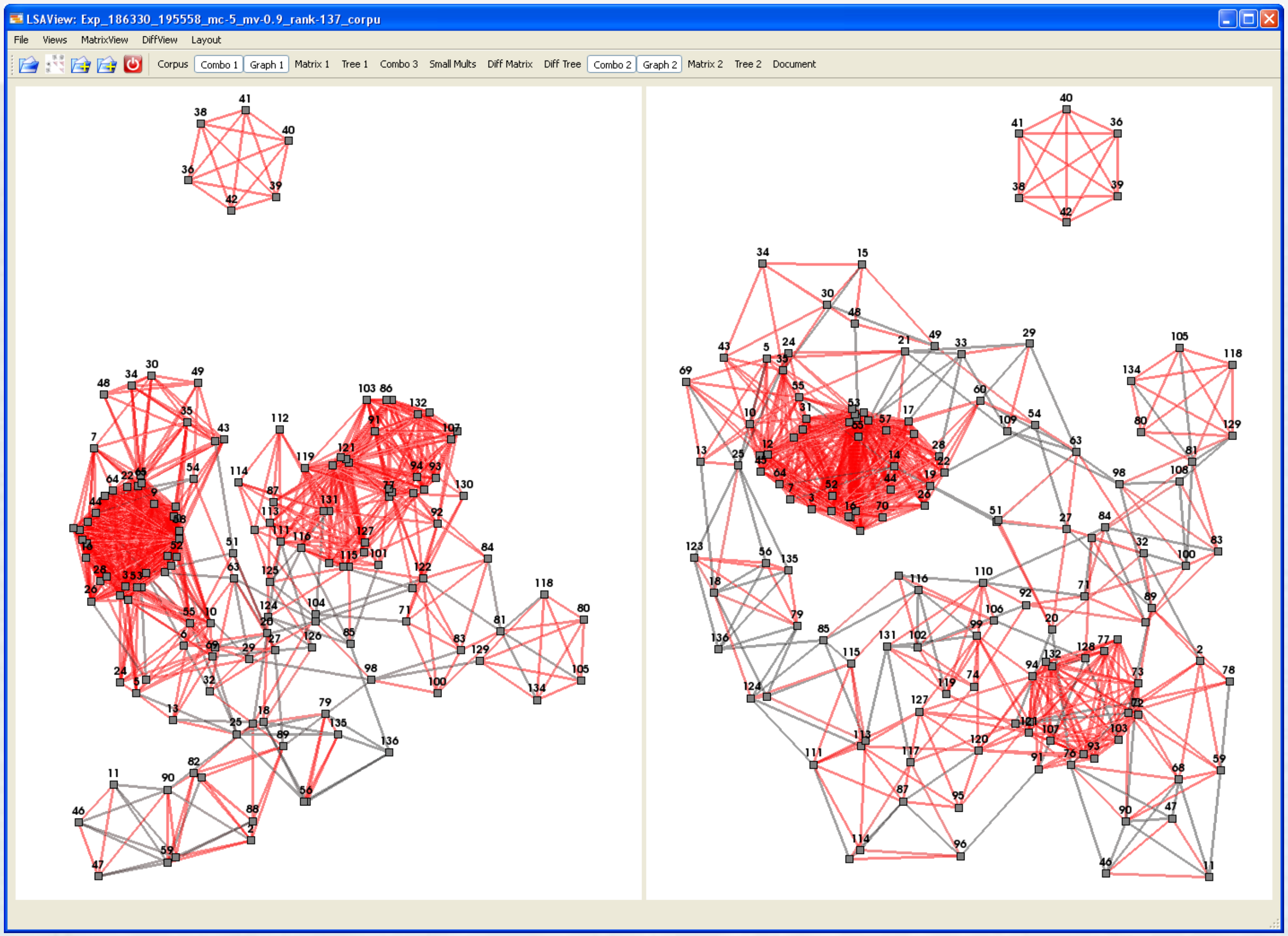




- Matrix views show edge weights
- k = 100 little difference in weights
- k = 20 good clustering

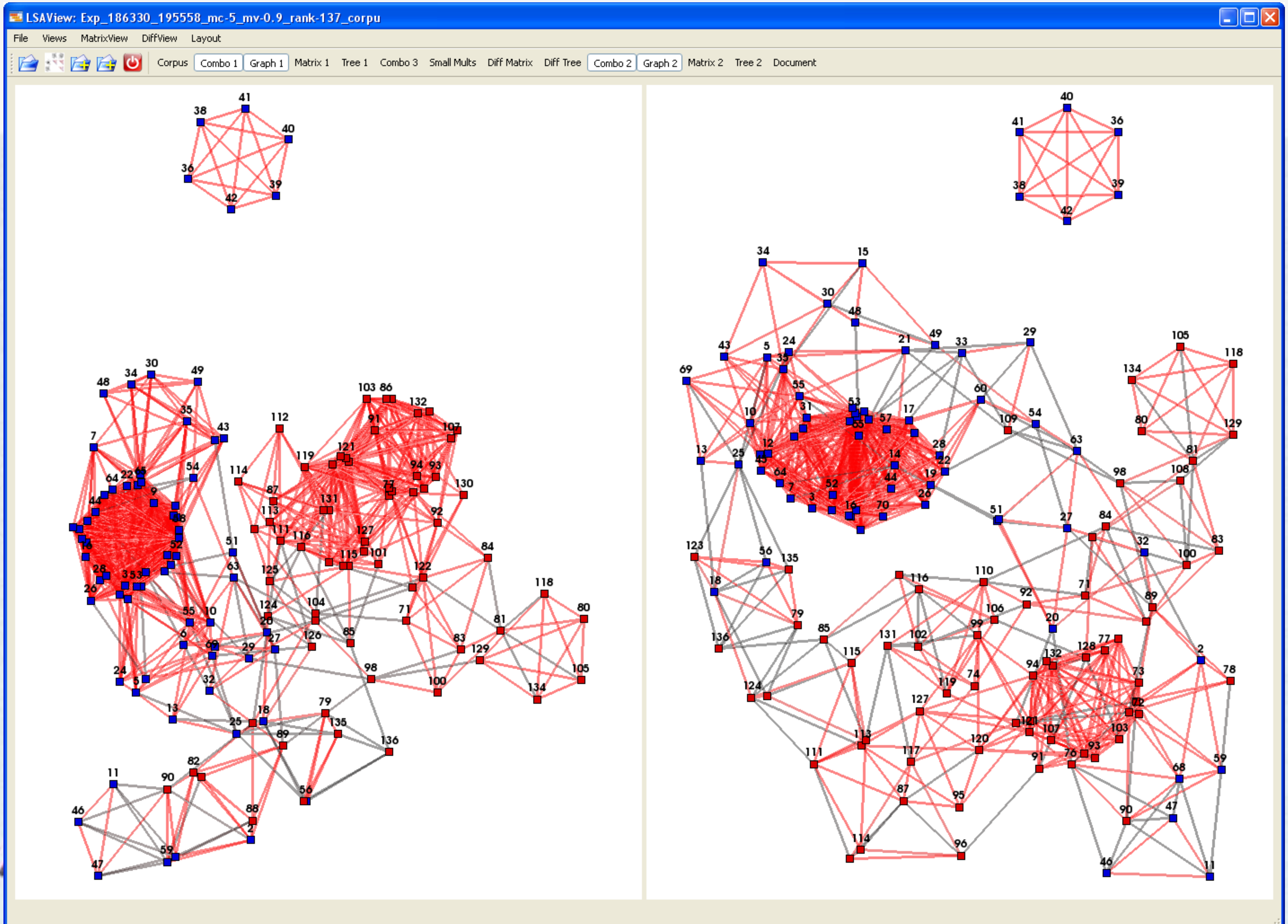


TECHTC k = 6, α = 1 vs. α = -1



- After further analysis, select k=6
- Both α have two distinct clusters
- Slightly stronger links in α = -1
- Both scalings perform well

TECHTC True Cluster Assignments



Conclusions

- Illustrated how LSAView used to understand LSA models
  - Seeding of other models (graph models)
  - Impact on document clustering task
- Key departure from previous work
  - Produces significantly different rank selection than automated approaches
  - Focuses on impact to text analysis tasks over variance

Work Funded by Laboratory Directed Research & Development (LDRD) program at Sandia National Laboratories

VAST 09

$$A_k = U_k \Sigma_k V_k^{\mathsf{T}}$$

$$e_{ij}(k) = \frac{\langle v_k^i \Sigma_k, v_k^j \Sigma_k \rangle}{\|v_k^i \Sigma_k\|_2 \, \|v_k^j \Sigma_k\|_2}$$

$$e_{ij}(k,\alpha) = \frac{\langle v_k^i \Sigma_k^{\alpha/2}, v_k^j \Sigma_k^{\alpha/2} \rangle}{\|v_k^i \Sigma_k^{\alpha/2}\|_2 \, \|v_k^j \Sigma_k^{\alpha/2}\|_2}$$

$$\bar{e}_{ij}(k,\alpha,n) = \frac{1}{n+1} \sum_{r=k-n/2}^{k+n/2} e_{ij}(r,\alpha)$$

$$s_{ij}(k,\alpha,n) = \sqrt{\frac{1}{n} \sum_{r=k-n/2}^{k+n/2} \left(e_{ij}(r,\alpha) - \bar{e}_{ij}(k,\alpha,n)\right)^2}$$

$$t_{ij}^{(1)} = \frac{\bar{e}_{ij}(k,\alpha,n) - e_{ij}(k,\alpha)}{s_{ij}(k,\alpha,n)/\sqrt{n+1}}$$

$$t_{ij}^{(2)} = \frac{\bar{e}_{ij}(k_1,\alpha,n_1) - \bar{e}_{ij}(k_2,\alpha,n_2)}{\sqrt{\frac{[s_{ij}(k_1,\alpha,n_1)]^2}{n_1} + \frac{[s_{ij}(k_2,\alpha,n_2)]^2}{n_2}}}$$